SUMMARIZED LINGUOGRAPHIC CORPUS OF THE TATAR LANGUAGE: ARCHITECTURE, STRUCTURAL PRINCIPLES

Karimullina Guzel Nurutdinovna, Karimullina Rezeda Nurutdinovna

Kazan (Volga Region) Federal University, 18 Kremlevskaya St., Kazan **(RUSSIA)**

DOI: 10.7813/jll.2015/6-2/24

Received: 14 Jan, 2015
Accepted: 16 Feb, 2015

ABSTRACT

The article gives general characteristic of a summarized linguographic corpus of the Tatar language with the sources including more then 360 various dictionaries. Some of them are represented by a considerable number of works (bilingual, terminological and other), a number of dictionaries are represented as an exclusive edition (dictionaries of abbreviations etc.), individual types of linguistic reference books have appeared quite recently (autonomous dictionaries of the writer's language).

Linguographic corpus of the Tatar language which is being created in Kazan University according to the Government Program "Preservation, study and development of the national languages of the Republic of Tatarstan and other languages of the Republic of Tatarstan for years 2014-2020" consists of 3 main components (modules): "Sources", "Linguographic characteristics (attributes, parameters)", "Vocabulary".

The "Sources" module provides a user with a variety of information about macrostructure of a dictionary, its target auditory, volume of vocabulary etc. The "Linguographic characteristics (attributes, parameters)" module is mainly aimed at provision of a user with information on content of the dictionaries as well as at demonstration of fullness of a parameter and the nature and methods of representation. The "Vocabulary" module is designed for combining and description of the vocabulary material of the source dictionaries of the Summarized corpus of the Tatar language dictionaries.

Information potential of the corpus allows usage of the presented materials at different stages of linguographic activity (design, composition, expert evaluation, editing of a linguistic reference book) as well as improvement of quality of the dictionaries being composed.

Key words: summarized corpus, linguography, dictionary, the Tatar language, information potential

1. INTRODUCTION

The last decades are characterized by rapid development of linguography which is an area of linguistics dealing with theory and practice of dictionaries (linguistic reference books) composition [1]. There have been and are being composed a lot of various types of dictionaries in the Tatar linguography which is due to the growing social demand for linguistic reference books [2, 3].

The modern science has a need in diversified interpretation of experience of theoretical and practical linguography, in a systematized analysis of the existing linguographic sources and of information presented in the dictionaries as well as in consequent elaboration of recommendations for the dictionaries improvement, in determining the ways and methods of elimination of deficiencies contained in the dictionaries.

The Tatar linguorgaphy offers various types of linguistic reference books [4, 5]. Some of them amount to considerable quantities: bilingual (Russian-Tatar and Tatar-Russian), terminological, phraseological, dictionaries of borrowings, orthological; a number of dictionaries have only one edition: inverse dictionary, dictionaries of abbreviations. Some types of dictionaries have appeared quite recently, namely autonomous dictionaries of the writer's language in the form of concordances were published no sooner then in the beginning of the XXI century.

Some types of the Tatar linguistic reference books are still not existent (for example historical, frequency dictionaries, associative dictionary, dictionary of paronyms etc.).

Different registering linguistic reference books which describe materials of the dictionaries are intended to play an important role in scientific and information support of a variety of areas of the Tatar and comparative linguistics inclusive of linguography.

By no means were all of the issues in the sphere of theory and specifically of practice of the Tatar linguography resolved to the full extent which undoubtedly affects the quality and usefulness of the dictionaries being prepared. In this respect the modern Tatar linguography requires systematized description of materials contained in the existing dictionaries.

Analysis of the current Tatar dictionaries has revealed that the vocabulary of a present-day Tatar standard language was not represented to the right degree (to the full extent). This makes reasonable actual need in creation of a corpus of materials from various dictionary sources.

Importance of this task is determined by the current status of the Tatar language as one of two official languages in the Republic of Tatarstan. The programs of implementation of the "Law on the languages of the peoples of the Republic of Tatarstan" state the necessity to enlarge the inventory of the Tatar language reference books and to compose the Tatar language dictionaries of various types.

2. LEXICOGRAPHIC CORPUS OF THE TATAR LANGUAGE DICTIONARIES: GENERAL CHARACTERISTIC

In order to solve the set task there has been taken the decision to create a Summarized Corpus of the Tatar language dictionaries which would describe all linguistic units which were documented in the Tatar language dictionaries. The present investigation is a part of a complex of projects included into the Government program "Preservation, study and development of the national languages of the Republic of Tatarstan and other languages of the Republic of Tatarstan for years 2014-2020".

The Summarized Corpus of the Tatar language dictionaries is designed for accumulation of data contained in the Tatar language dictionaries of the second half of the XX and the beginning of the XXI century and is being developed as a multifunctional informational linguographic system which arranges and structures dictionary data in a special way allowing manipulations with the mentioned data (or their components), systematization, comparison and selection of material as consistent with the set tasks.

The Summarized Corpus of the Tatar language dictionaries consists of three main information modules which are the system components having some inherent value:

- the "Sources" module,
- the "Linguographic characteristics (attributes, parameters)" module,
- the "Vocabulary" module.

3. THE "SOURCES" MODULE

The "Sources" module provides a user with a variety of information about macrostructure of a dictionary, its target auditory, volume of vocabulary etc.

This stage involves all of the general dictionaries of the second half of the XX and the beginning of the XXI century which reflect the materials of the Tatar standard language (except for terminological, dialectological, encyclopedic, onomastic and students' dictionaries) and which are mentioned in the bibliographic reference book "Tatar linguography: dictionaries of years 1951-2008" (Kazan, 2011) [6] as linguographic sources for the Summarized Corpus of the Tatar language dictionaries.

The above reference book is one of the main components of the "Sources" module of the Summarized Corpus of dictionaries. A computerized version gives an opportunity to supplement it quickly and effectively with new information zones and data on the Tatar language dictionaries already published or being published.

Every dictionary is described according to a special scheme consisting of the following paragraphs:

- Complete bibliographic description with an indication of all authors, number of pages; in case of bilingual titles or two title pages (in the Russian and the Tatar languages) the description with use of the source language comes first.
- Bibliographic data (ISBN, UDC, LBC, author mark, dictionary size in printed sheets, dictionary format, number of printed copies)
- Synopsis.
- Dictionary structure.
- Size of vocabulary.
- Dictionary entry samples (as a rule a fragment of the fist page of the dictionary vocabulary is shown).

The "Sources" module of the Summarized Corpus of the Tatar language dictionaries allows to analyze (select) the linguistic reference books based on various attributes; for example:

by the number of languages – mono-, bi-, multilingual,

by language orientation – monoscopal, biscopal,

by selection of units – general and special etc.

In the Tatar linguistics classification of dictionaries by types is scarcely represented which prevents the Tatar linguistic reference books from being provided with sufficiently full and adequate characteristic and description. We are not aware of existence of special papers dedicated to typology of the Tatar linguistic reference books. Usually the review of the types of dictionaries was given in the Tatar language textbooks, bibliographic indices [7, 8, 9] and individual Tatar dictionaries. In quite considerable number of cases segregation of different types of linguistic reference books has extremely single-sided nature, many dictionaries may be classified as belonging to different types but are described in the context of one and the same type.

For example, in the existing dictionary indices a) the "Explanatory dictionaries" section includes not only the Tatar language explanatory dictionary but also "Explanatory dictionary and reference book in physics, chemistry and chemical engineering in the Russian, Tatar and English languages", however the similar linguistic reference books (for instance, "Pedagogy explanatory dictionary", "Shorter Russian-Tatar explanatory dictionary of medical terms (with equivalents in the English, German, French and Latin languages)" etc.) are mentioned in the "Terminological dictionaries" section; b) the "Terminological dictionaries" section includes onomasticons: a dictionary of hydronyms, a dictionary of microtoponyms, a dictionary of personal names, a dictionary of word-formative elements; c) "Pocket Tatar-Russian and Russian-Tatar dictionary" was included to the "Pocket dictionaries" section but was not mentioned in the "Bilingual dictionaries" section; d) the dictionaries of hydronyms and mycrotoponyms were included to the "Thematic dictionaries" section while the dictionary of personal names was not referred to any of the groups.

4. THE "LINGUOGRAPHIC CHARACTERISTICS (ATTRIBUTES, PARAMETERS)" MODULE

The principal task of the "Linguographic characteristics (attributes, parameters)" module is to provide a user with information which is contained in the dictionaries, to demonstrate fullness of a parameter and the nature and methods of representation.

In the Summarized Corpus of the Tatar language dictionaries a user can get a description of the methods of expression of every parameter as well as select the necessary dictionary and obtain its parametric representation (a list of the parameters documented in it and the methods of representation). It is important to note that this component, i.e. parametric characteristic of a linguistic reference book, is absent in bibliographic indices. Based on the parametric representation one can understand what kind of information is contained in this dictionary, what are the methods of its representation. The parametric representation is accompanied by illustrative dictionary entries in which the areas of realization of a particular parameter are outlined.

"Parameter" as a linguistic term for the first time was considered by Yu.N. Karaulov in his work "Linguistic construction and formal language thesaurus" [10]. In the course of analysis we used a system of parameters developed for the lynguographic corpus and allowing description of various types of information contained in the dictionaries (for more details see [11]). The system includes 25 parameters:

ANTONYMS [AN]. Indication of antonyms of a word.

MEANING [ME]. Textual description of a word meaning, availability of a definition.

INFLEXION [IN]. Inflexion characteristic of a word.

ILLUSTRATION [IL]. Visual image explaining a word semantics.
WORD ETYMOLOGY [WE]. A word historical and etymological information.
FOREIGN-LANGUAGE EQUIVALENT [FE]. Translation to other languages, foreign-language equivalents.
CULTURE-ORIENTED LINGUISTICS [COL]. Linguistic and cultural information about a word, short description of culture-specific concepts.
POLYSEMY [PS]. Indication of a word polysemy.
MORPHOLOGY [MPH]. Morphologic characteristic of a word.
MORPHEMIC DIVISION [MD]. Specification of a morphemic composition of a word.
SPELLING [SP]. Written image of a word.
HOMONYMS [HOM]. Indication of homonyms of a word.
SPECIFIC NATURE OF USE [SU]. Stylistic, expressive characteristic of a word, sphere of use.
PARONYMS [PS]. Indication of paronyms of a word.
PROVERBS AND SAYINGS [PRO] Indication of paremiaes where a title unit is used.
PRONUNCIATION [PR]. Audio presentation of a word which is communicated through a transcription.
EXAMPLES OF USE [EU]. A phrase, a sentence where a title unit is used.
SYNONYMS [SY]. Indication of synonyms of a word.
WORD-FORMATION [WF]. Data on a word which was used in formation of a title unit, on words formed with use of the TU, on conjugate words with regard to the TU.
SYNTACTIC CHARACTERISTIC [SCH]. Explains syntactic properties of a word.
THEMA [THE]. Thematic characteristic of a word.
ACCENT [ACC]. Information on place and nature of a word accent.
USAGE [US]. Information of frequency of use, usage of a word.
PHRASEOLOGY [PHR]. Indication of phraseological units where a title unit is used.
PART OF SPEECH [PoS]. Specification of a word belonging to a definite part of speech.
Such classification is based on the following principles:
1. A parameter should characterize a linguographic unit but not structural features of a dictionary.
2. A parameter should characterize a linguographic unit and a unit of language with due account for its linguistic peculiarities.

Each parameter can be described in more detail given its structure and specific supplementary data. In addition the parameters have various methods of representation in the dictionaries (metalinguistic means). For example the parameter of meaning (ME) is being represented through 1) a foreign-language equivalent, 2) an interpretation, 3) an explanation and a supplement, 4) a picture, 5) illustrative examples; the morphologic parameter (MPH) for different parts of speech is being communicated in a number of ways: the substantives may be accompanied with specification of a gender, the verbs with specification of an aspect/ transitivity/non-transitivity, the numerals and the pronouns with specification of a category etc.

In a dictionary a parameter may be given for all units (an "A" label in a table of parameters) of for a group of units (a "G" label in a table of parameters).

Parameter	AN	ME	IN	IL	WE	FE	COL	PS	MPH	MD	SP	HOM	SU
04-TP$_{(88...04)}$↑. OO: words (title units -TU), phrases (like ***бетмә!***, rarely); word combinations (all units - AU), phraseological units (PHU (AU)		A$^{\perp}$	G			A$^{\perp}$		G	G$^{\perp}$		A	G	G
	PA	**PRO**	**PR**	**EU**	**SY**	**WF**	**SCH**	**THE**	**ACC**	**US**	**PHR**	**PoS**	
							G		A		G	A	
	ME – explained by means of a foreign-language equivalent; explanations are given for a part of units (see ***әбый,*** 2nd mean.); labels "*fig.*" and "*dir.*" are given for the units with direct or figurative meaning (see ***абынырга***); **IN** – case forms are given for the pronouns (see ***мин***); **FE** – if a unit is translated with use of a different part of speech (noun, adj., adv.) it is provided with the sign // (see ***абзац***); **SP** – variants of spelling are given for some words (see ***ашамсак***); **SU** – for individual units stylistic, terminological or special labels are given like *folk.*, *myth.*, *colloq.* (see ***әдһм,*** 2nd mean.; ***әдһми***) etc.; **SCH** – for impersonal verbs *imper.* label is specified (see ***љшетергә***, 2nd mean.); for individual units the translated (Russian) part contains declension questions (see ***бизергә***)); **ACC** – is given for all words in Russian; accent for Tatar words is shown when the accent falls on any syllable except the last one (see ***һмма***); **PHR** – is given within a dictionary entry after ◊ sign (see ***агач***); **PoS** – direct specification of a part of speech.												

Information on parameters gives an opportunity to find out what kind of data on a dictionary unit, to what extent and of what type is given in the dictionaries as well as to make observations and conclusions as to informative fullness and quality of the existing Tatar linguographic sources.

Thus for example the ME parameter is a principal parameter for explanatory dictionaries. In the course of analysis it was revealed that this parameter is not equally represented within one dictionary. For instance in [04-TP$_{(88...04)}$↑ (hereinafter the indices from the dictionaries contained in the bibliographic reference book "Tatar linguography: dictionaries of 1951-2008" (Kazan, 2011) [6]) are given]: **доллар** <...> dollar (*monetary unit*), but compare to **франк** <...> franc; **әлүчә** <...> cherry plum (*a tree and a fruit*), **миләүшә** <...> violet (*a flower*) <...>, but compare to **кольраби** <...> kohlrabi, **кыртым** <...> safflower, **лиана** <...> liana etc.

As a rule the HOM parameter is represented in monolingual explanatory and unabridged bilingual dictionaries. This parameter is the principal one for the dictionaries of homonyms. In the course of analysis of this parameter there were revealed some deficiencies and non-conformities contained in the dictionaries. For example in "Tatar-Russian and Russian-Tatar school dictionary: (for Russian-speaking students)" [08-TP_PT_шк.] a number of units are presented as monosemantic or polysemantic however in "School dictionary of the Tatar language omonyms" [97-TP_омон_шк_*Сафин*.] the same author mentions them as homonyms.

In the linguographic sources synonyms (the SY parameter) are represented as referenced entries, for example: in [98-TP_антропон_*Сат.] –* ***Бүләк*** *<...>.* Синонимнары: *Гата, Зайд, Нәфил <...>;* ***Кашкарбай*** *<...>.* Синонимнары: *Бүребай, Чанбай.*

The words having close meaning are marked with compare *label after which the unit with the meaning close to the meaning of the word is stated, for instance:* ***капкыч*** *<...>;*.чагышт.: *мәм <...>;* ***чолдор*** *<...>;* чагышт.: *лапы I, 2. <...>.*

5. THE "VOCABULARY" MODULE

The "Vocabulary" module is designed for combining and description of the vocabulary material of the source dictionaries of the Summarized Corpus of Tatar language dictionaries.

A summarized vocabulary is a basis of this module. The summarized vocabulary of the modern Tatar standard language will allow obtaining miscellaneous information on a word: accentologic, morphologic (of a part of speech), data on the type of representation of homonymic and polysemantic units in various dictionaries.

The "Vocabulary" module may provide miscellaneous information on composition of vocabulary of the sources in the form of quantitative characteristics or a list of words:

- In regard of one dictionary – а) all words documented in a dictionary inclusive of an alphabetical listing of the sources units where the words are arranged not according to the alphabet: b) title units of main and referenced dictionary entries; c) intra-entry words (all of them or some of their types in accordance with the characteristics shown in the Summarized Corpus of the Tatar language dictionaries); d) homonyms; e) words represented in a dictionary in other then the root form; f) words having any formal signs for example definite letters, combination of letters, morphemes etc.
- In regard of two or more dictionaries (in addition to the data specified in the paragraph above) – а) unified vocabulary; б) words common for the sources under consideration; в) lexical units documented in one source (sources) and absent in other source (sources).

The module materials give an opportunity to reveal errors, omissions within the dictionaries content, to provide recommendations for the compilers and experts of the linguographic reference books.

The materials of the corresponding linguographic classes are not always taken into account in the course of selecting a vocabulary for a linguistic reference book, sometimes allowance is made for the minimum two-component linguographic classes (for example of antonymic nature); more details are given in [5, 11].

The results of comparative analysis of the dictionaries within the Summarized Corpus show that the dictionary "Let's speak Tatar: Tatar-Russian and Russian-Tatar dictionary" [98-ТР_РТ_*Хар.*] contains ***алтын*** but doesn't contain ***кълмеш,*** contains ***аю*** but doesn't contain ***тълке, куян,*** contains ***гасыр*** but doesn't contain ***ел,*** contains ***кньп, тиз, яхшы*** but doesn't contain ***ђз, ђкрен, начар.*** The book "Basic Tatar-Russian and Russian-Tatar dictionary for secondary school students" [08-ТР_РТ_шк_*Саф*$_{(97...08)}$↑] in the Tatar-Russian part has ***тълнлђ, клндез, кич белђн***, but does not have ***иртђн;*** the names of only two months (***октябрь, гыйнвар***) were given, the rest 10 are absent.

Quite often derivative words in the dictionaries are given without the commonly used initial lexical units; namely "Tatar-Russian, Russian-Tatar school dictionary" [08-ТР_РТ_шк.] contains ***бозлы*** but does not contain ***боз;*** contains ***велосипедчы, трактрчы, футболчы, фокусчы*** but does not contain ***велосипед, трактор, футбол, фокус,*** contains ***егермелђп*** but does not contain ***егерме*** unit etc.

The results of data comparison from the Tatar language dictionaries revealed omissions of various types of units, fro example [05-Т_толк] does not contain: ***гөнаһсыз*** (contained in 66-ТР, 07-ТР_1,2), ***җырчы*** (contained in 77-Т_толк$_{(77-81)}$, 07-ТР_1,2), ***йөзләп*** (contained in 77-Т_толк$_{(77-81)}$, 66-ТР), ***калу*** (contained in 77-Т_толк$_{(77-81)}$, 66-ТР, 07-ТР_1,2), ***моңсу*** (contained in 77-Т_толк$_{(77-81)}$, 07-ТР_1,2), ***сүнмәс*** (contained in 66-ТР, 07-ТР_1,2), ***унлап*** (contained in 77-Т_толк$_{(77-81)}$, 07-ТР_1,2), ***учлап*** (contained in 66-ТР, 07-ТР_1,2), ***эчеру*** (contained in 66-ТР, 07-ТР_1,2) etc.

Therefore the summarized dictionary will allow comparison of its vocabulary with vocabulary of the source dictionaries, revelation of omissions in the dictionaries which is of high significance for compiling of vocabularies for the dictionaries being in the process of preparation and for supplementing and improvement of the republished Tatar language dictionaries.

The materials represented in the summarized vocabulary of the Tatar language dictionaries will provide information for analysis and investigation of a wide range of phenomena and cases.

6. CONCLUSION

Comprehensive registration and consolidated description of materials of the sources within the Summarized Corpus of the Tatar language dictionaries will provide a user with an opportunity to obtain information about structure and content of the linguistic reference books, about the degree of representation of definite units in the same, to reveal omissions in the linguographic sources as well as to get data on specific characteristics of such units, their system interconnections (accentology, a part of speech, polysemy, homonymy, synonymy, antonymy).

High scientific and information potential of the Summarized Corpus, ample opportunities of work with data included in it make it a multifunctional and effective tool for creation and expert evaluation of various types of dictionaries and a basic source for linguographic activity at the different stages of the same.

CONFLICT OF INTERESTS

The author confirms that the presented data do not contain any conflict of interests.

ACKNOWLEDGEMENT

The work is performed according to the "Action Plan for implementation of the Program of Competitive Growth of the Federal State Autonomous Educational Institution of Higher Professional Education "Kazan (Volga Region) Federal University" among the leading world research and educational centers for years 2013-2020".

REFERENCES

1. Computer linguography / scientific editorship. N.K. Zamov, K.R. Galiullin. – Kazan: Publishing House of Kazan University, 1995.– 119 p. (Online edition.: http://old.kpfu.ru/f10/publications/1995/K1.php).
2. Galiullin K., Gizatullina A., Gorobets E., Karimullina G., Karimullina R., Martyanov D.: Corpus-Based Regiolect Studies: Kazan Region. In: Lecture Notes in Computer Science, pp. 169-175, vol. 8773. Speech and Computer (2014)
3. Galiullin, K., Gorobets, E., Karimullina, G., Karimullina, R.: Computational Corpus of Tatar Proverbs and Sayings: Electronic Database of Paremias. In: Phraseology in Multilingual Society, pp. 350–362. Cambridge Scholars Publishing, Newcastle upon Tyne (2014)

4. Yusupova A.Sh.: Tatar Language Dictionaries of XIX Century as a Unified.-World Applied Sciences Journal 30 (2): 186-190 (2014)
5. Karimullina R.N. Bilingual Tatar linguography of the second half of the XX – the beginning of the XXI century: thesis of ... Cand. Sc. (Philology)/ R.N. Karimullina.- Kazan, 2011.- 253 p.
6. Tatar linguography: dictionaries of 1951-2008: bibliographic reference book / compiler R.N. Karimullina; scientific editorship of K.R. Galiullin.- Kazan: Kazan University, 2011.- 528 p.
7. Minullin K. Татар теле – сњзлеклђрдђ: библиографик күрсәткеч = Tatar language in dictionaries: bibliographic index [1801-1998] / K. Minnullin, R. Valiullin; scientific editorship of A.R. Rakhimov.- Kazan: Master Line, 1998.- 56 б.
8. Хәкимҗан Ф.С. Татар тел белеме библиографиясе (1981-1997) / Ф.С.Хәкимҗан, Т.Х.Хәйретдинова.- Казан: Татар дәүләт гуманитар ин-ты нәшр., 1998.- 204 б.
9. Yakupova G.K. Bibliographic reference for Tatar linguistics (1778-1980) / G. Yakupova; editorial staff.: Sh.N. Asylgaraev, L.T. Makhmutova (executive editor), T.Kh. Khairutdinova.- Kazan: Tatar Book Publishing House, 1988.- 136 c.
10. Karaulov Yu.N. Linguistic construction and formal language thesaurus / Yu.N. Karaulov.- M.: Nauka, 1981.- 366 p.
11. Galiullin K.R. Russian and Tatar linguography: scientific and information support: thesis of ... Cand. Sc. (Philology) / K.R. Galiullin.- Kazan, 2000.- 343 p.